RNAI POTENCY PREDICTION METHOD

Background of the Invention

The present invention relates generally to the field of RNA interference (RNAi) and particularly to methods for obtaining RNAi reagents of high RNAi potency using an algorithm trained by an artificial neuronal network.

Field of the Invention

Since Elbashir et al. have demonstrated the ability of synthetic small interfering RNA (siRNA) to mediate, through RNA interference (RNAi) mechanism, specific mRNA down regulation in mammalian cells [see, for instance, Elbashir et al., Nature, Vol. 411, pp. 494-498 (2001); and Caplen et al., PNAS, Vol. 98, No. 17, pp. 9742-9747 (2001)], this technique has been increasingly used as a research tool to investigate gene function by exploring the phenotypes induced by the specific down-regulation of a given gene. In particular, as gene silencing experiment by mean of an siRNA or RNA-interference type of reagent requires only the knowledge of a part of the nucleotide sequence of the targeted gene, it is envisageable, once the genome of a particular organism is known, to design for every gene a RNAi experiment and screen for phenotypes relevant for instance for therapeutic targets discovery. This genome-wide approach dictates the specifications for the design of the RNAi reagents. Indeed, specificity parameters avoiding off-target silencing become of particular importance. Also, the characterization of individual siRNAs for potency is no longer envisageable. Consequently, potency prediction algorithms are required for the application of RNA gene silencing experiment to genome-wide phenotypic screens.

It has been suggested that similarly to antisense, target accessibility plays an important role in the potency of siRNA. See Kretschmer-Kazemi et al., Nucleic Acids Res., Vol. 31, No. 15, pp. 4417-4424 (2003). Recently another study [see Anastasia Khvorova, Cell, Vol. 115, pp. 209-216 (2003)] revealed some sequence requirements triggering siRNA and miRNA potency such as notably a low internally stability of the dsRNA at 5' of the antisense strand as well as in the region 9-14 of the antisense strand. These findings were obtained by a statistical analysis of the relationship between siRNA potency and duplex internal

thermodynamic stabilities. The study was based on siRNA potencies of 375 randomly selected siRNA against three different targets.

Given the fact that the RNAi mechanism is not fully characterized and that many additional parameters may have an impact on siRNA potency, it would be beneficial to acquire larger functional data sets in order to have a better understanding of the sequence-activity relationship.

Summary of the Invention

In one aspect, the present invention pertains to a method of making an algorithm for the prediction of the RNAi potency of a RNAi reagent comprising:

a) determining experimentally the potency of a plurality of RNAi reagent to down regulate a reporter protein readout; and

b) using said potency data set to train a artificial neuronal network.

In another aspect, the present invention pertains to an algorithm obtained by a method according to the present invention.

In another aspect, the present invention pertains to a method for predicting the RNAi potency of a RNAi reagent comprising:

a) providing a plurality of RNAi reagent sequences comprising a region complementary to a given target gene;

b) applying a trained artificial neuronal network according to the present invention to said RNAi reagent sequences; and

c) selecting the RNAi reagent sequence(s) which are predicted to be potent.

In another aspect, the present invention pertains to a method for inhibiting the expression of a given target gene, comprising:

a) providing a plurality of RNAi reagent sequences comprising a region complementary to a given target gene;

b) applying a trained artificial neuronal network according to the present invention to a said RNAi reagent sequences;

c) selecting the RNAi reagent sequence(s) which are predicted to be potent;

d) synthesizing the RNAi reagent(s) selected in c);

e) exposing cells expressing the target gene with the RNAi reagent(s) of d); and

f) measuring the activity of the RNAi reagent or other phenotypes which may be induced by the down regulation of the target gene.

Description of the Figures

Figure 1: An example of a normalized data set. 79 siRNA targeting a 3'-UTR insert of a YFP mRNA were co-transfected with the reporter fusion mRNA (H1299, 50 nM, readout at 50h). Grey bars are 3'-UTR targeting siRNA, black bars are positive and negative controls. Negative control was arbitrary set at 10% potency and positive control at 90% potency. Each siRNA potencies were normalized according to these controls.

Figure 2: Illustrates the filtering of screening data.

Figure 3: Comparison of prediction versus screening measures for the training set.

Figure 4: Comparison of prediction versus screening measures for the testing set.

Figure 5: Dependence of the prediction – measure correlation (on test set) from the size of the training set.

Detailed Description of the Invention

All patent applications, patents and literature references cited herein are hereby incorporated by reference in their entirety.

As used herein, the terms "RNAi reagent" and "oligoribonucleotide" are used interchangeably and mean an oligomer or polymer of ribonucleic acid (RNA) or deoxyribonucleic acid (DNA) or a mimetic thereof. The RNAi reagent may also comprise modified ribonucleotide residues. Suitable modifications are known in the art. See, for instance, Uhlmann, Current Opin. Drug Discovery Dev., Vol. 3, No. 2, pp. 203-213 (2000); and Uhlmann and Peyman, Chem. Rev., Washington, DC, Vol. 90, No. 4, pp. 543-584 (1990). The term RNAi reagent encompasses both single-stranded and double-stranded nucleic acid molecules. Double-stranded nucleic acid molecules may be composed of two

separate strands or of one strand comprising two regions which can form a double-stranded structure and a spacer region between the two regions forming a hairpin loop. In the context of RNAi, the RNAi reagents are preferably of double-stranded structure and comprise a sequence complementary to a target gene. However, the present invention is not limited to double stranded structures and includes also single-stranded RNAi reagents capable of inducing RNAi. See Schwarz et al., Mol. Cell, Vol. 10, No. 3, pp. 537-548 (2002).

In a first aspect, the present invention provides a method of making an algorithm for the prediction of the RNAi potency of a RNAi reagent comprising:

a) determining experimentally the potency of a plurality of RNAi reagent to down regulate a reporter protein readout; and

b) using said potency data set to train a artificial neuronal network.

In one embodiment, the method of making an algorithm for the prediction of the RNAi potency of a RNAi reagent comprises the steps of:

a) determining experimentally the RNAi potency of a plurality of RNAi reagents comprising a sequence complementary to at least one target gene;

b) generating a data set of the potency of said RNAi reagents with the experimentally determined RNAi potency of a), wherein said data set though obtained and different targets (reporter fusion-mRNA) has a common protein read-out which is normalizable by mean reporter specific positive and negative controls; and

c) training an artificial neuronal net using said read-out.

In the context of the present invention, the term algorithm means a set of equations and a set of rules which can be applied to the data automatically and can be implemented as a code which executes on a computer.

"RNAi potency" or "potency" is a term of the art and means the relative ability of a given siRNA to down-regulate a given protein or mRNA upon its transfection in a cellular assay. Typically, potency of a siRNA is determined by measuring the level of expression of the target mRNA or protein and is usually expressed as a percentage of the negative control. Thus, a high potency means that the RNAi reagent is capable of efficiently inhibiting, i.e., decreasing, the expression of a target gene, whereas a low potency means that the expression of the target gene is not or only little inhibited. A potent RNAi reagent inhibits the

expression of a target gene more than 50% as compared to negative control, preferably more than 60%, more than 70%, more than 80%, most preferably more than 90%.

The RNAi reagent in accordance with the present invention is a RNAi reagent suitable for RNAi experimentation. Various types of RNAi reagents suitable for RNAi are known in the art. See Dykxhoorn et al., Nature Rev., Vol. 4, pp. 457-467 (2003). Such a RNAi reagent comprises a sequence that is complementary to a target gene. Complementary to a target gene within the context of the present invention means that the sequence is complementary to an RNA transcribed from the DNA sequence of the target gene, including pre-mRNA, mRNA, cDNA. The term "target gene" is meant to include any DNA sequence which is expressed, i.e., transcribed to RNA, in a cell, tissue or organism. The expressed sequence needs not necessarily be translated into a protein and includes for instance also pre-mRNAs, regulatory RNAs, rRNAs, etc. The sequence complementary to the target gene typically is from about 19-23 nucleotides long, but may also be longer. Preferably, the complementary sequence is less than 50 nucleotides long, more preferably from 15-35 nucleotides long or from 18-25 nucleotides long. The complementary sequence is preferably 100% identical to the corresponding sequence of the target gene, i.e., there is no mismatch between the complementary sequence and the corresponding sequence of the target gene. In some embodiments, the complementary sequence may comprise one, two, three, four, five or more mismatches, if these mismatches do not abolish the RNAi activity of the RNAi reagent. The RNAi reagent used for RNAi is preferably double-stranded and may be composed of two separate strands, but may also be composed of one strand forming a hairpin loop. The double-strand RNA region of the RNAi reagent may contain mismatches and function in a miRNA like mechanism. Examples for types of RNAi reagents that mediate RNAi are for instance siRNAs or miRNA (microRNA) or small hairpin RNAs (shRNAs).

In a one step, the RNA potency of a plurality of RNAi reagents is experimentally determined in accordance with the methods of the present invention. In general, it is advantageous to provide a large number of RNAi reagents, as there is a positive correlation between the number of RNAi reagents that are provided and the quality of the algorithm. However, for practical reasons as RNAi reagent synthesis and the experimental determination of RNAi potency is costly and time-consuming, it may be desirable to keep the number of RNAi reagents for experimental determination of RNA potency as low as possible. However, the number of RNAi reagents should not be below a minimum number, below which the

algorithm is not properly trainable. In a preferred embodiment, at least 10 RNAi reagents, at least 50 RNAi reagents or at least 100 RNAi reagents are provided, in a more preferred embodiment at least 200 RNAi reagents, at least 500 RNAi reagents, at least 1000 RNAi reagents or at least 2000 RNAi reagents are provided. In another preferred embodiment, less than 10000 RNAi reagents, preferably less than 5000 RNAi reagents or more preferably less than 3000 RNAi reagents are provided. In another preferred embodiment, the RNAi reagents are randomly selected. The RNAi reagents may be overlapping or not overlapping, in a preferred embodiment, the RNAi reagents are not overlapping. The RNAi reagents comprise a region complementary to a target gene. Several RNAi reagents may comprise a region complementary to the same target gene, overlapping or not overlapping. In one particular embodiment, all RNAi reagents comprise a region complementary to the same target gene, overlapping or not overlapping. In another embodiment, the RNAi reagents comprise a region complementary to more than one target gene, preferably to at least 2 target genes, at least 5 target genes or at least 10 target genes.

In another preferred embodiment of the present invention, the RNAi reagent sequences provided in step a) were pre-screened for RNAi specificity. "RNAi specificity" or "specificity" is a term of the art and refers in the context of the present invention to the selectivity of a RNAi reagent, i.e., to the capability of a RNAi reagent to selectively inhibit or decrease the expression of a given target gene without inhibiting or decreasing the expression of other genes expressed in a cell, tissue or organism. Ideally, specific RNAi reagents only inhibit the expression of the target gene and leave the expression of all other genes expressed in a cell, tissue or organism unaffected. For this purpose, the specific RNAi reagents advantageously comprise a fully complementary sequence to the target gene, i.e., a complementary sequence with no mismatch to the target sequence, but no fully complementary sequence to all other genes expressed in a cell or an organism, i.e., said complementary sequences have at least one mismatch, preferably at least two mismatches, more preferably at least three mismatches to the all sequences expressed in a cell, tissue or organism other than the target gene. Pre-selection of the specificity of the RNAi reagents can for instance be done by computationally comparing the sequence of the RNAi reagents of interest to all known expressed sequences available in the databases for a given cell, tissue or organism using a suitable software for sequence comparison.

Numerous methods suitable for experimentally determining RNAi potency of a RNAi reagent are known in the art. Generally, a double-stranded RNAi reagent comprising a complementary region to a given target gene is transfected into a cell expressing the target gene. For transfection, different methods, such as for instance electroporation, use of cationic lipids or cationic polymers as helpers for transfection may be used. The cells are then incubated under suitable conditions allowing the expression of the target gene. The expression of the target gene is subsequently measured using a suitable technique, such as, for instance, RT-PCR or measuring the amount of a reporter protein. In a preferred embodiment, a fusion-mRNA coding for a reporter is co-transfected with the siRNA. Preferably, the target nucleotide sequence is inserted in 3'-UTR of the reporter coding sequence. As such, the target will not be translated and its down regulation will have no biological impact. In a preferred embodiment the reporter protein is a Yellow Fluorescent protein (YFP). In another embodiment, negative and reporter specific controls, i.e., a siRNA targeting the coding region of the reporter protein and thereby silencing the reporter protein with a similar potency independently of the 3'-UTR insert, are used and allow the comparison of the expression level of the reporter protein obtained with each siRNA with the expression levels of the negative control and the reporter specific positive control. As such expression levels measured for all of the siRNAs can be compared and pooled in one single homogeneous data set. Basically, any kind of cell may be used for transfection, however, in a preferred embodiment, the cell is a eukaryotic cell, preferably an animal cell, more preferably a mammalian cell and most preferably a human cell.

Upon comparison of the inhibition observed for each siRNA with the positive and negative controls (also mentioned as normalization), an experimental potency score correlating to the experimentally determined RNAi potency will be created for each RNAi reagent leading to a read-out which can be compiled in a data set of the experiments. In a preferred embodiment, the experimental scores are obtained by measurement of the reporter protein. Preferably all data of the experimental read-out come from a single type of experimental setting under homogenous conditions. The data is preferably based on a measurement on protein level and not mRNA level, i.e., the amount of protein expressed by the target gene is measured upon exposure with the RNAi reagent and not the amount of mRNA. A typical experimental setup for experimentally determining the RNAi potency in accordance with the present invention is described hereinbelow in the Examples.

In a preferred embodiment of the present invention, a reporter assay is used for the experimental determination of RNAi activity. The use of a reporter assay in accordance with the present invention allows to screen a large number of siRNA against a broad panel of targets with a common experimental read-out. Such an assay is described in Hüsken et al., Nucleic Acids Res., Vol. 31, No. 17, p. e102 (2003). Briefly, an mRNA fusion transcript construct comprising of a full-length reporter gene with a target region of interest inserted into the 3'-untranslated region is provided. For instance, Luciferase and Fluorescent reporter genes can be used in the constructs. The RNAi reagents to be tested for RNAi potency comprise a sequence complementary of interest inserted into the 3'-untranslated region. The RNAi reagent is then transfected with a suitable transfection method into cells expressing transiently or constitutively the reporter gene construct and cultured under suitable conditions allowing the expression of the reporter gene. The level of protein expressed by the reporter gene is subsequently measured. Such an assay allows measuring the RNAi potency of large numbers of RNAi reagents at the protein level against a broad panel of targets with a common read-out. As such, the generated data set is homogeneous and all potency data are comparable each together.

The above described data set is used for training of an artificial neuronal network. Artificial neuronal networks are known in the art, see, for instance, Zell, Simulation neuronaler Netzwerke, Addison Wesley (1994); and Rumelhart and McLelland, Parallel Distributed Processing, Vol.1, MIT Press, Cambridge, MA (1986), and available, for instance, on http://www-ra.informatik.uni-tuebingen.de/SNNS. Statistical information will be extracted from the siRNA sequence antisense strand by artificial neuronal networks and correlated with the screening measures. Eventually, a trained network can be applied to an arbitrary input sequence to give an estimate of the 'would-be' screening measure. For instance, for the purpose of illustration, a 3 layer feed-forward network with back propagation training 7-8 can be used. The input layer consists of 4 lanes of ordered nodes, see Figure 1. There is one lane per nucleic base type and always 4 nodes of different base type referring to the same input sequence position. The number of positions is the length of the input sequence. At any given time, during training and/or application, exactly one node at any given position is activated. The activities along the ordered lanes then represent the input sequence. The signals of activated nodes are propagated from the input layer into the second layer, also called layer of hidden units. During this propagation the signals of the input layer, either 0 or

1, are weighted differently, summed into the signals of the hidden units. Similarly the signals of the hidden units progress into a single output node of the third and last layer. The weights are the memory elements to represent the statistical knowledge. Initially the weights are randomly set, and result in output signals for siRNA antisense sequence which deviate from the true screening signals. The difference between current network output signal and the experimental is used to change all weights to reduce the differences. Back propagation comes into place by the network for having a 'true' target signal for hidden units in the second layer.

Another aspect of the present invention provides a computer system comprising computer hardware and the algorithm of the present invention. A further aspect of the present invention provides a computer readable medium comprising the algorithm of the present invention.

Another aspect of the present invention provides methods for obtaining RNAi reagents with an enhanced likelihood of RNAi potency against a given target gene, i.e., an enhanced likelihood of inhibiting the expression of a pre-selected target gene. Thus, for a given number of RNAi reagents designed, a higher percentage will be potent using this method to design RNAi reagents than if the RNAi reagents were designed randomly. Conversely, fewer RNAi reagents have to be designed and screened to find a given number of RNAi reagents of high RNAi potency useful for specifically inhibiting the expression of a target gene. In one embodiment, the method according to the present invention comprises the steps of:

a) providing a plurality of RNAi reagent sequences comprising a region complementary to a given target gene;

b) applying the trained algorithm according to the present invention to said RNAi reagent sequences using a neuronal network; and

c) selecting the RNAi reagent sequence(s) which are predicted to be potent.

In a first step, a plurality of candidate RNAi reagents sequences comprising a sequence complementary to the gene to be targeted, i.e. to the gene to be inhibited by RNAi, is chosen. The RNAi reagents may be prescreened for specificity or for the presence or absence of certain sequence motives. They may be overlapping or non-overlapping. In one embodiment, the candidate RNAi reagents are randomly chosen. In a second step, the

potency of the RNAi reagents provided in the first step is predicted using the neuronal network which is trained with an algorithm in accordance with the present invention. In a next step, the RNAi reagent sequences which are predicted to be potent are selected. For instance, the three or five or 10 most potent RNAi reagents can be chosen. Alternatively, all RNAi reagent sequences with a prediction score above a certain threshold score are chosen. In a preferred embodiment, the threshold score is at least 0.7, at least 0.75, at least 0.8 or at least 0.85. The selected RNAi reagents can now be experimentally assayed for their RNAi potency. Thus, in a next step, RNAi reagents suitable for RNAi comprising the sequences which were predicted to be active are synthesized. In a preferred embodiment, the RNAi reagents are chemically synthesized. The skilled person is familiar with chemical methods for the synthesis of such oligonucleotides, for instance, through the well-known technique of solid phase synthesis. However, the RNAi reagents may also be synthesized using biochemical methods such as *in vitro* transcription or vector based systems. Suitable cells expressing the target gene can now be exposed to the synthesized RNAi reagents (or to the vector comprising the sequence of interest in case of a vector based system), incubated under suitable conditions and the expression level of the target gene can be measured with suitable methods. As a control, the expression level of the target gene of cells which were not exposed to the sequence of interest can be compared.

The following examples are included to demonstrate preferred embodiments of the invention and are not intended to limit the invention.

EXAMPLES

In the approach illustrating this invention more than 3,000 siRNA targeting 34 different mRNA have been screened for potency in a cellular assay. One feature of this study has been to generate a homogeneous data set for subsequent analysis of potency-sequence relationship. This has been possible by the use of a fusion-mRNA reporter assay. See Hüsken et al. (2003), *supra*. In this assay, a plasmid encoding a reporter fusion-mRNA for a reporter protein where the target sequence has been inserted in 3'-UTR of the reporter mRNA is transfected followed by the transfection of the RNAi reagent. As such:

a) the result of down regulation of the target sequence has no biological consequences;

b) in all assays, the potency readout is at protein level and as the same reporter protein is the whole study, the potency data are not biased by different read-outs; and

c) the use of common positive and negative controls in all assays allows the normalization of all potency data.

A artificial neuronal network has been used to investigate the sequence potency relationship of this homogeneous potency data se. As a result, the trained artificial neuronal network is able to predict the potency of any siRNA solely based on its nucleotide sequence. As the sequence requirements for potency of gene silencing reagents acting through RNA interference pathway, this approach will be applicable to other RNAi reagents like, for instance, shRNAs or miRNAs.

RNAi Reagents

The RNAi reagents used in that study were 21-mer double-stranded RNAi reagents with a 19 base-pairing RNA region and dideoxynucleotide overhangs on 3' of each strand. The overhang of the sense strand was systematically a dithymidine whereas the overhangs of the antisense was a dideoxynucleotide designed to be complementary to the target.

Screening Methods: eYFP mRNA-Fusion Reporter Assay

Construction of reporter expression clones

The enhanced cyan and yellow fluorescent protein (eCFP, eYFP) dual reporter based vector pNAS-092 (described in Hüsken et al. (2003), *supra*), was constructed to contain a multiple cloning site after the stop codon of the eYFP for inserting the appropriate cDNAs or ESTs of interest. The eCFP reporter serves for normalization measurements driven under the elongation factor 1 alpha (EF-1α promoter and the eYFP reporter was used to monitor activity of siRNA driven under the CMV promoter. The origin of the vector was a plasmid pBudCE4 (Invitrogen) which contains a hCMV and a EF-1α promoter. pNAS-092 was generated by inserting the eCFP gene derived from peCFP-N1 (Clontech) and by transferring the eYFP gene derived from peFP-N1 (Clontech) together with a synthetic DNA fragment bearing the cloning site (EcoRV, NotI, HindIII, KpnI, XbaI). The synthetic DNA that was used in the pNAS-092 was confirmed by sequencing. For alternative cloning strategies pNAS-092 was converted to a GatewayTM destination vector pNAS-097 by inserting the

attR1 and attR2 cloning sites according to the manufacture protocol (Invitrogen) after the eYFP stop codon. All plasmids used for the final reporter assay were constructed by inserting into the cloning site the c-DNA via ligation (pNAS-092) or recombination (pNAS-097).

Cell Lines and Cell Culture

The human non-small cell lung carcinoma cell line H-1299 (CRL-5803) was purchased from ATCC (Rockville, MD). H-1299 cells was maintained in 5% humidified CO_2 atmosphere at 37°C in RPMI 1640 medium (Life Technologies) containing 10% fetal bovine serum plus 1% L-Glutamine. H1299 cells were split 48 hours prior to transfection reaching a 80% subconfluent stage. Cells were trypsinized, washed and equally dispensed (50 µL) in black 96-well assay plates (Costar, clear bottom) a day before the transfection.

Fluorescent protein assays with dual reporter construct carrying a reference gene (intracellular normalization).

Plasmid Transfection

Lipofectamine-PLUS reagent was incubated with the plasmid diluted in OptiMEM-I (22 ng/µL plasmid, 4.4 mL/mL Lipofectamine-PLUS) and then this mix was diluted 11-fold with OptiMEM-I. Lipofectamine was 1.3-fold pre-diluted with HEPES (20 mM, pH 7.2) and diluted further 28.6-fold with OptiMEM-I (26.6 µL/mL Lipofectamine) and kept for 15 minutes. Both mixtures were combined 1:1 and incubated for 15 minutes, further diluted 10-fold with OptiMEM-I. The medium was aspirated and 100 µL was added to the cells (0.2 µL/mL Lipofectamine-PLUS, 13.3 µL/mL Lipofectamine, 1 ng/µL plasmid). After 2 hours, 50 µL siRNA transfection mixture was added to the cells which was then further incubated for 2 hours.

siRNA Transfection

Oligofectamin was diluted with OptiMEM-I (60 µL/mL) were mixed and incubated for 30 minutes at room temperature. The siRNA was diluted to 600 µM from the hybridized stock solution with hybridization buffer (30 mM HEPES, 100 mM Potassium Acetate, 2 mM Magnesium Acetate: pH 7.63 at room temperature. Annealing 2 minutes at 90°C followed by 1 hour at 37°C). Diluted Oligofectamin and siRNA were combined 2 volumes plus

1 volume, and incubated for 15 minutes. The siRNA-Oligofectamin mix was further diluted 1:1 with OptiMEM-I and transferred onto the cells (50 µL) (final conc. 0.7 ng/µL plasmid, 10 µL/mL Oligofectamin, 50 nM siRNA). The medium was removed and replaced with 100 µL standard RPMI medium without phenol red containing 10% fetal bovine serum plus 1% L-Glutamine and incubated for 3 days at 37°C. Fluorescence was measured in 24-hour intervals. The fluorescence of eCFP and eYFP was measured with the excitation filter of 436/20 nm and the emission filter of 480/30 nm and the excitation filter of 500/25 nm and the emission filter of 535/30 nm, respectively. The quotient of eYFP/eCFP fluorescence counts expresses the eYFP activity per cell number equivalent. For this data collection with the eYFP reporter assay all siRNA treatments of the positive standard (YFP-specific siRNA NAS-12842/58) and negative standard (luciferase siRNA NAS-8548/9) were done in triplicate. The standard deviation from the mean of the standard siRNAs NAS-12842/58 treatments was calculated and found on average to be 9.1%.

Target Selection

Firstly, reporter plasmids with 34 different inserts were selected. Size of the inserts varied between 344 nucleotides and 3784 nucleotides.

siRNA Sequence Design

With 79 siRNA per plate we have 3160 in total. The sequences were designed to walk the inserts randomly, allowing for overlap of varying size (0-20 bases). With an insert size of 27k bases even regularly chosen positions of the 3160 siRNA would result in significant overlap (13 bases) in the siRNA sequences. Sequences with long polynucleotide stretch (five or more consecutive nucleotides) were not considered. In cases of long inserts, two sets of 79 siRNA were designed.

The total set of 3160 siRNA sequences was checked for its nucleic content. It was found that all possible motifs up to tetra nucleotides were present in the siRNA sequence set.

Screening Format

Each siRNA plate contains 79 siRNA, a negative control (anti-luciferase siRNA NAS-8549), two reporter specific siRNA (anti-YFP siRNA NAS-12842 and NAS-12847). Control siRNA

were pipetted in from one single batch in triplicates. Eight wells were left empty and will be used for "plasmid only" negative controls.

Filtering and Normalization of siRNA Activity Data (see Figure 2)

Each plate was assayed in duplicates and YFP level was measured by fluorimetry at two time points. For each plate, linear correlation within the duplicate was checked, as well as the level of inhibition of the positive and negative controls. Data were accepted when a correlation within the duplicate was superior to 0.7 and when the positive control was down regulating YFP to at least 60% as compared to negative control (see Figure 3). Data set from five assays were rejected according to that filter. Remaining data set (2717 sequences) was divided in a training set and a testing set and further filtered by checking each siRNA individually and excluding siRNA displaying a variation higher than 30% within the duplicate. As such, approximately 15% of data were further removed. The remaining duplicate data points were averaged to result in a data set with reduced noise characteristics. The final data set contained 2109 sequences in training set and 234 sequences in testing set. All data points S(i), i indexing the data points, were normalized by the affine system A

$$T(i) = A(S(i)) = (T\_high-T\_low) / (S\_high-S\_low) * (S(i)-S\_low) + T\_low,$$

in which the original signal of the negative control which is S_low and shall be normalized to give T_low (we set T_low = 0.1 = 10%). Similarly defined is the positive control signal S_high and its transformed T_high (set to 0.9 = 90%).

Artificial Neuronal Network Training (see Figure 3)

The siRNA sequence data is presented to the input layer and the screening reporter signals are used to adjust the weights between the network nodes. Each siRNA sequence and its screening measure is presented a total of 10 times. After a one time presentation of all data points the weights of the network are updated synchronously with a learning rate of 0.1 and a momentum of 0.1. See Zell, Simulation Neuronaler Netzwerke, Addison Wesley (1994). Based on five different initializations of weights, the resulting five trained networks weights differed but all five networks showed consistently only slightly varying prediction output. A final output was achieved by averaging the signals of the respective output node of all five

networks. We will refer to the average output as 'the' output of 'the' network instead of looking at properties of any of the individual networks for simplicity.

Evaluation of Performances of Predictor (see Figure 4)

The trained network predicted the experimental inhibition activity with a correlation of 0.63 when applied to the testing set, while it showed moderately higher correlation of 0.665 when applied to the training set. Figure 3 graphs the concordance of the two. In addition to the correlation between prediction and experimental values, the performance of the algorithm can be evaluated by setting a threshold for experimentally active siRNA and for predicted active siRNA. The threshold for experimentally active siRNA was set at 75% normalized potency (0.75). The threshold for predicted active siRNA was set at a score above 0.8. These thresholds form four quadrants containing true negatives (predicted inactive and inactive), false negative (predicted inactive but active), false positive (predicted active but inactive) and true positive (predicted active and active) sequences. The predictor performance can be measured by its sensitivity and selectivity.

Predictor sensitivity = true positive / (true positive + false negative) = 0.26

Predictor selectivity = true positive / (true positive + false positive) = 0.71

These values indicate that the predicted sequences have a 71% probability to be active (as defined above). This value has to be compared to the hit rate observed on the whole testing set where 35% of the sequences were active. Also, the predictor will identify 26% of the active sequences.

Influence of the Size of the Training Set on the Predictor's Performance (see Figure 5)

The training data does not need to be taken completely for training, which allows investigating the degradation of the BIOpred prediction performance with fading set size. The size of the testing set was constant. The correlation, see Figure 5, consistently degrades slowly with reduced training sets. The correlation for a training data set as small as 265 data points still gives correlations of about 0.53.